AD_____

Award Number: W81XWH-06-1-0127

TITLE: A Strategy to Rapidly Re-Sequence the NF1 Genomic Loci Using Microarrays and Bioinformatics for Molecular Classification of the Disease

PRINCIPAL INVESTIGATOR: Sitharthan Kamalakaran
Josh Dubnau

CONTRACTING ORGANIZATION: Cold Spring Harbor Laboratory
Cold Spring Harbor, NY 11724

REPORT DATE: December 2006

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT	3. DATES COVERED
01-12-2006	Final	17 Nov 2005 – 16 Nov 2006

4. TITLE AND SUBTITLE

A Strategy to Rapidly Re-Sequence the NF1 Genomic Loci Using Microarrays and Bioinformatics for Molecular Classification of the Disease

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-06-1-0127

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Sitharthan Kamalakaran
Josh Dubnau

Email: dubnau@cshl.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Cold Spring Harbor Laboratory
Cold Spring Harbor, NY 11724

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
Original contains colored plates: ALL DTIC reproductions will be in black and white.

14. ABSTRACT

Neurofibromatoses (NF) are disorders caused by mutations in NF1/NF2 genes and characterized by fibromatous tumors on nerves, skin, and bones. A method to rapidly and cheaply re-sequence the NF loci would greatly aid in the molecular classification of the disease by forging links between sequence variations and clinical manifestations.We propose to develop a technique to rapidly re-sequence genomic loci using microarray hybridization and bioinformatics. The NF locus is first amplified using a high processivity polymerase and hybridized on a custom microarray containing all possible 10mer combinations of the four deoxy-ribonucleotides. A virtual profile of a hybridization map of the locus is generated using the available sequence information. This map is then compared to an actual hybridization. If the sequences are identical, then the two images would superimpose. Using the differences in the virtual and real hybridization, we propose to map the mutations in the locus.

15. SUBJECT TERMS
Not provided

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	8	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

Introduction:

Neurofibromatoses (NF) are disorders caused by mutations in the NF1 or NF2 gene loci and characterized by fibromatous tumors on nerves, skin and bones. However, the clinical manifestations of NF are extremely variable, and can also include optic gliomas, scoliosis, hypertension and learning disabilities. The genetic basis for the variability in clinical manifestations is not understood. Molecular characterization of the gene loci in multiple patients would greatly aid classification of the disease. The size of NF genomic loci (280kb and 115kb respectively), however, together with practical limitations of sequencing technologies preclude such studies. Thus a method to rapidly and cheaply re-sequence the NF loci in a high throughput manner would greatly aid in forging links between gene sequence and clinical manifestations.

Body:

We proposed to develop a technique to rapidly re-sequence large genomic loci using microarray hybridization and bioinformatics. Starting with the known sequence of the NF1 locus on Chr17, it is possible to bioinformatically design multiple degenerate primer sets specifically for this region. Using a linear isothermal genome amplification method using the φ29 polymerase (which is higly processive and capable of amplifying stretches of 10kb), we proposed to amplify a genomic representation of this locus with high specificity[1, 2]. This NF1 locus "amplicon" then would be labeled and hybridized on a custom chip containing all possible 10mer combinations of the four deoxy-ribo nucleotides, which would constitute 1048576 probes. This probe number can easily be arrayed with current technologies[3]. A saturated sampling of the amplified region using the high density arrays would have enough information to identify sequence information for hundreds of kilobases of DNA. This is done by creating a virtual profile of a hybridization map of the locus. This map is then compared to an actual hybridization image. If the sequences are identical, then the two images would superimpose. However if there are point mutations, single nucleotide polymorphisms (SNPs) or deletions in the DNA sample, the image map would be shifted accordingly. For example, consider a 15 base region AGTCTTA***G***GATCCGA and its SNP/mutation AGTCTTA***C***GATCCGA. This single base change will shift the hybridization of 20 probes for the 10 base sequences from 9 bases before and after the altered base (from AGTCTTA***G***GA to AGTCTTA***C***GA, GTCTTA***G***GAT to GTCTTA***C***GAT and so forth). Our interest in this approach also was bolstered by the realization that such a 10mer chip would theoretically be universally applicable to re-sequencing any gene region from any species. The 'specificity' would derive from the amplicon chosen for hybridization.

The "wet lab" methodologies for differentiating single base changes have already been optimized for SNP genotyping with microarrays. Using intensity maps of observed and expected hybridization values for a given locus, one can devise an optimization algorithm capable of reconstructing the most likely alterations with high specificity and sensitivity.

Our informatics manipulations of micro-array hybridization data using oligo arrays with suggested to us a practical drawback of this method caused by the high level of cross-hybridization obtained with short oligos. Even with 20-mer oligos, this issue would likely preclude the success of our strategy. The existing microarray technologies do not allow for more than 2 million probes to be spotted on a single array. While, this is enough to spot all combinations of DNA sequences on length 10 only (4^10 = 1 048 576) or one half of all combinations of DNA sequences of length 11 (4^11 = 4 194 304), 10 probes cannot be used to detect exact hybridizations. We performed several experiments to measure the extent of this cross-hybridization noise that comes from microarray hybridizations. The most convincing example comes from tiling array data that we received from Dr Robert Lucito (CSHL) from hybridizations that were performed on a Nimblegen chip. The probes on this array are 50 bases in length and are from the human genome. In order to collect the noise from a microarray, we used a hybridization of the array to a labeled sample from a female tissue. This permitted us to estimate the cross-hybridization noise from the intensity of hybridization signals of the probes from Y chromosome. We plotted signal intensities of probes from Y chromosome and compared it with signal from Chr10 as control. Fig 1 shows a plot of the frequency of probes against intensity of the signal. As can be seen from Fig1, there are a number of Y chromosome probes that report as well as those from Chromosome 10.

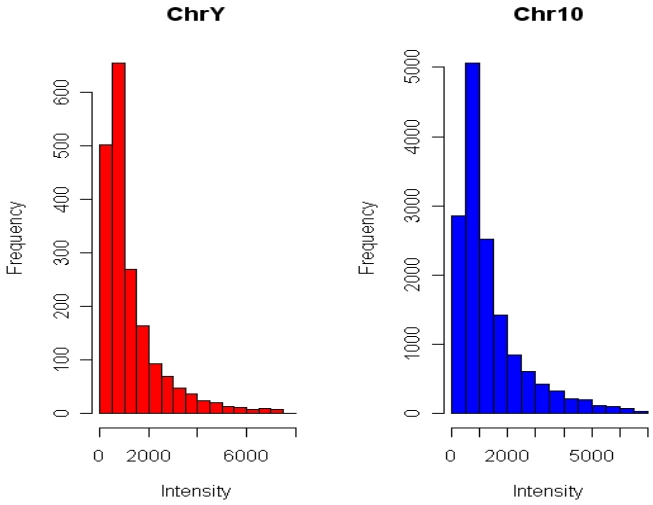

Fig 1. Signal intensity bins are on the X axis and the number of probes in each bin is plotted as a histogram.

While our proposed strategy will tolerate some cross hybridization, we realized that the use of 10-mers will not be feasible. For our approach to succeed, we would need to increase the oligo length and the stringency of the hybridization. Based on results from current array hybridizations, we determined that the sequence length of the probe sequences must be at least 25 bases.If we were to array all possible sequences of length 25, that would amount to $1.12589991 \times 10^{15}$ probes. This would be not feasible given today's technology.

We next attempted to reduce the complexity be limiting to 25-mer sequences that are present in the human genome. In order to most efficiently array all sequences of length 25 in the genome on the array we decided to explore the use of De Bruijn Sequences[4]. In combinatorial mathematics, a k-ary De Bruijn sequence B(k, n) of order n is a cyclic sequence of a given alphabet A with size k for which every possible subsequence of length n in A is present exactly once.
Such a sequence has the following properties:

* Each B(k, n) has length kn
* There are k!^{k^{(n - 1)}}/k^n distinct De Bruijn sequences B(k, n).

Thus if we consider the DNA to be an alphabet of size 4 (A/T/C/G), we can determine the most efficient way of representing all the words of length, 'k'. We fix an alphabet with four characters. Fix a positive integer k. Let \Sigma be the set of all words in the alphabet of length k. We call a circular sequence \sigma a _de Bruijn_ sequence if every element of \Sigma appears exactly once as a substring of \sigma.

If \sigma and \tau are de Bruijn sequences call them _opposing_ sequences if they do not share any common subsequence of length k+1. Now we can ask the following questions:
1. Do opposing sequences exist?
2. Is there a decent way of producing them, given k?
3. Are there any 'canonical' opposing pairs?

Call \sigma and \tau _m-opposing_ if for any pair of substrings of length m, say \sigma' in \sigma and \tau' in \tau, there is at most one word of length k appearing in both \sigma' and \tau'. Now the question is there a way of producing them given k and m? In a 1993 paper, Robert Rowley and Bella Bose proved the existence of fairly large families of pairwise opposing sequences, which can be used to construct parallel opposing DeBruijn sequences[5, 6]. This algorithm can be effectively used to construct multiple opposing DeBruijn Sequences which can then be arrayed on a microarray. Having multiple probes that target the same sequence of DNA will allow one to find different regions of the genome that are hybridizing to the array. However the problem becomes more complicated for one to decipher if there are any variances in the intensity of hybridization that occur purely because of the kinetics and chemistry of the hybridization and not due to varying concentrations of the DNA being measured.

Exploration of the kinetics of the hybridization led us to the second major problem. Deciphering data from microarray hybridizations is hugely complicated by the variance in intensity of probes that measure the same concentrations of DNA, but whose sequence varies. In the microarray field, this problem presents a huge limitation that doesn't allow the two probes of different sequence identities to be compared effectively. Quantitative detection of DNA requires that microarray probes exhibit a sensitive and predictable response to concentrations of specific targets of the probes. This response must occur in the presence of a complex mixture of nonspecific targets[7]. This presents an additional problem that needs to be overcome for our method to work. We performed a

hybridization of whole genome drosophila DNA onto a tiling microarray from Affymetrix. This array has a 25b probe tiled along the genome of Drosophila with at least one probe every 35b of the genome. This hybridization allowed us to evaluate the differences between hybridization intensity across probes from the same region of DNA. The concentration of DNA from a single region will be the same. However the difference in probe to probe hybridization intensity is very high (Fig 2).

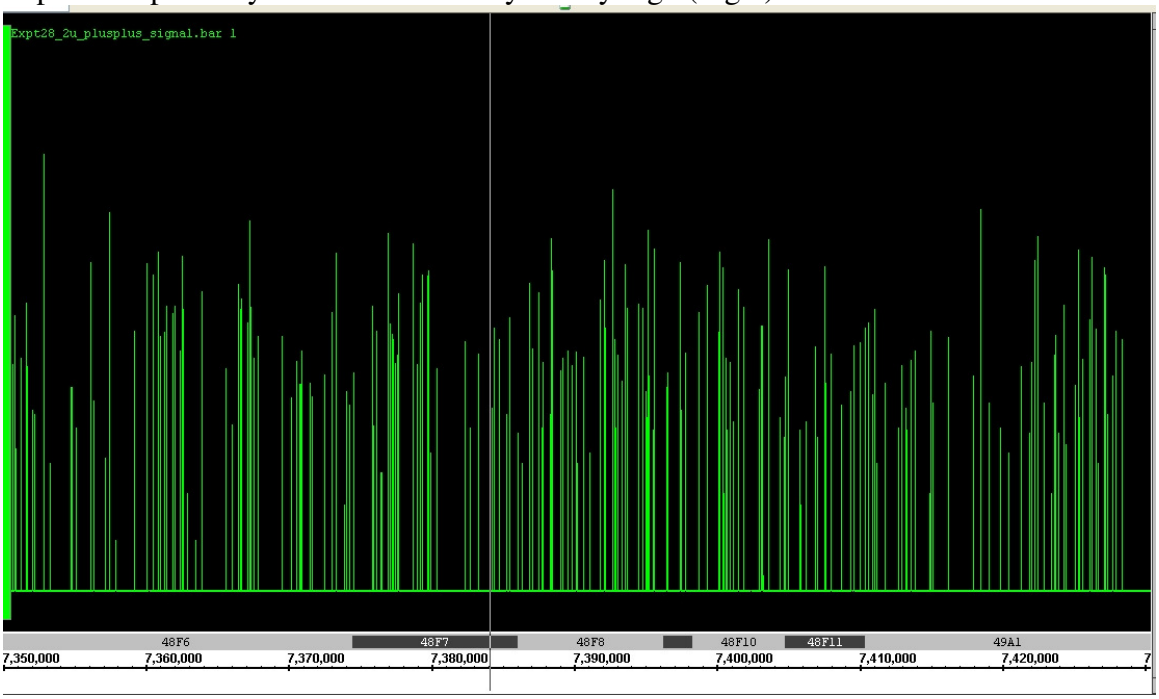

Fig2. Hybridization intensity along a 75kb region of Drosophila genome. Each line in the graph is a signal intensity

In light of these problems, we decided to abandon this method of resequencing using a universal oligo microarray. Such arrays while theoretically possible are too limited by the kinetics of hybridization for one to reliably resequence a segment of DNA.

KEY RESEARCH ACCOMPLISHMENTS:

1. The feasibility of using microarrays for sequencing was studied. It was determined that microarrays have numerous technical problems that preclude their use in the way proposed
2. Theoretical methods were explored by which a given set of DNA sequences can be spotted most efficiently on a microarray by use of Debruijn graphs.

REPORTABLE OUTCOMES:

None

CONCLUSION:

We explored the use of microarrays for sequencing large regions of the genome. However the limitations of accurately identifying sequence specific hybridization with short probes (10bp probes) as envisioned was not feasible. Also the large variations in hybridization intensities for DNA that are of the same concentration but different sequence identity presented insurmountable problems in accurately deciphering DNA sequence from microarray hybridization. With the advent of many other sequencing technologies such as 454 sequencing and Solexa sequencing, it was determined that our approach was not the best approach for a cost effective and high throughput sequencing technology. In light of these problems, we decided to abandon this method of resequencing using a universal oligo microarray. Such arrays while theoretically possible are too limited by the kinetics of hybridization for one to reliably resequence a segment of DNA.

REFERENCES:

1. Dean, F.B., et al., *Comprehensive human genome amplification using multiple displacement amplification.* Proc Natl Acad Sci U S A, 2002. **99**(8): p. 5261-6.
2. Lasken, R.S. and M. Egholm, *Whole genome amplification: abundant supplies of DNA from precious samples or clinical specimens.* Trends Biotechnol, 2003. **21**(12): p. 531-5.
3. Bertone, P., et al., *Global identification of human transcribed sequences with genome tiling arrays.* Science, 2004. **306**(5705): p. 2242-6.
4. de Bruijn, N.G., *A Combinatorial Problem.* Koninklijke Nederlandse Akademie v. Wetenschappen 1946. **49**: p. 758–764.
5. R.A. Rowley, B.B., *On the Number of Arc-Disjoint Hamiltonian Circuits in the de Bruijn Graph.* Parallel Processing Letters, 1993. **3**(No. 12): p. 375-380.
6. R.A. Rowley, B.B., *Fault-Tolerant Ring Embedding in de Bruijn Networks* IEEE Transactions on Computers, 1993. **42**(12): p. 1480-1486.
7. Mei, R., et al., *Probe selection for high-density oligonucleotide arrays.* Proc Natl Acad Sci U S A, 2003. **100**(20): p. 11237-42.